

# Coevolution of trustful buyers and cooperative sellers in the trust game

Naoki Masuda<sup>1,2,\*</sup> and Mitsuhiro Nakamura<sup>1</sup>

**1** Department of Mathematical Informatics, The University of Tokyo, Bunkyo, Tokyo, Japan

**2** PRESTO, Japan Science and Technology Agency, Kawaguchi, Saitama, Japan

\* E-mail: masuda@mist.i.u-tokyo.ac.jp

## Abstract

Many online marketplaces enjoy great success. Buyers and sellers in successful markets carry out cooperative transactions even if they do not know each other in advance and a moral hazard exists. An indispensable component that enables cooperation in such social dilemma situations is the reputation system. Under the reputation system, a buyer can avoid transacting with a seller with a bad reputation. A transaction in online marketplaces is better modeled by the trust game than other social dilemma games, including the donation game and the prisoner's dilemma. In addition, most individuals participate mostly as buyers or sellers; each individual does not play the two roles with equal probability. Although the reputation mechanism is known to be able to remove the moral hazard in games with asymmetric roles, competition between different strategies and population dynamics of such a game are not sufficiently understood. On the other hand, existing models of reputation-based cooperation, also known as indirect reciprocity, are based on the symmetric donation game. We analyze the trust game with two fixed roles, where trustees (i.e., sellers) but not investors (i.e., buyers) possess reputation scores. We study the equilibria and the replicator dynamics of the game. We show that the reputation mechanism enables cooperation between unacquainted buyers and sellers under fairly generous conditions, even when such a cooperative equilibrium coexists with an asocial equilibrium in which buyers do not buy and sellers cheat. In addition, we show that not many buyers may care about the seller's reputation under cooperative equilibrium. Buyers' trusting behavior and sellers' reputation-driven cooperative behavior coevolve to alleviate the social dilemma.

## Introduction

The number of transactions executed in online marketplaces such as eBay is soaring up. To buy a desired item, a buyer must first trust in a seller by paying in advance. Because the seller may lose little by dismissing a single buyer, the seller may be tempted to ship a counterfeit item or may not even transport the purchase to the buyer. If many sellers behave maliciously toward buyers, the marketplace would collapse. Here is a moral hazard. A classical example in which counterfeit items would prevail is the “market for lemons” or the used car market [1]. Nevertheless, many auction sites and related online services, including opinion forums, price comparison sites, and product review sites, enjoy prosperity without seriously being swamped by the malicious behavior of users [2–5].

The main mechanism to elicit the cooperative behavior of sellers in such a social dilemma situation is the online reputation management system, also called the feedback mechanism [2–5]. When a reputation management system is implemented, a buyer is invited to evaluate the seller after a transaction so that other buyers can refer to the reputation of this seller in the future. A seller with a good overall reputation would successfully sell items to many buyers in the long run, whereas a seller with a bad reputation would be avoided by buyers. In a seminal paper, Klein and Leffler analyzed the role of reputations in alleviating a moral hazard [6].

Reputation mechanisms in online marketplaces are often cited as a successful example of indirect reciprocity [7,8]. Indirect reciprocity (precisely, downstream or vicarious reciprocity as its major subtype [8,9]) is a mechanism for the alleviation of social dilemmas. It dictates that an individual  $i$  with a good reputation is helped by another individual that  $i$  has not met.  $i$  may also help other individuals that  $i$  does not know. In fact, most buyer-seller pairs conduct only one transaction on eBay, such that a buyer probably does not know the seller with whom the buyer is about to transact [10]. It is theoretically established that indirect reciprocity enables cooperation in social dilemma games under proper conditions [8,9,11–16].

However, the mechanism governing cooperation between unacquainted buyers and sellers observed in real online transactions does not resemble that provided by these models. The existing models of indirect reciprocity are mostly based on the donation game, which is a type of social dilemma games. In the donation game, two players are chosen from a population, one as donor and the other as recipient. If the donor helps the recipient, the recipient gains a benefit, which is larger than the cost that the donor is

charged. If the donor does not help the recipient, the donor and the recipient gain nothing. The donor's help contributes to social welfare, whereas the donor is better off withholding the help.

In the previous models of indirect reciprocity [8, 9, 11–22], each player is selected as many times as donor and recipient per generation (we discuss an exception [23] in the Discussion). Therefore, the players are essentially involved in a symmetric prisoner's dilemma game. In contrast, the social dilemma game effectively played in online marketplaces is a highly asymmetric game. Buyers and sellers are distinct roles that each individual does not play with equal probability [10]. In addition, in the donation game and the symmetric prisoner's dilemma, there is no concept wherein one player invests trust in a peer player in the one-shot game. Theoretical models of reputation-based cooperation using different social dilemma games such as the ultimatum game also assume symmetrization of the two roles [24, 25]. This is also the case for the models of reputation-based cooperation analyzed in economic literature [26, 27].

The trust game [28–31] seems to be a much better model for online marketplaces [7]. As shown in Fig. 1, the buyer, also called the investor, first decides whether to buy an item from the seller. If the buyer buys, the seller, also called the trustee, decides whether to ship the appropriate item to the buyer. If the buyer buys and the seller does not ship, succumbing to the temptation to defect, the seller gains the largest payoff 1, and the buyer gains the smallest payoff  $-1$ . If the buyer buys and the seller ships, both players obtain a relatively large payoff  $r$ , where  $0 < r < 1$ . If the buyer does not buy, both players gain a relatively small payoff 0, which is the unique Nash equilibrium of the game. A cooperative society is realized when the buyer buys and the seller ships. In laboratory experiments, humans cooperate to some extent in the trust game [30–33], and reputation mechanisms enhance cooperation [7, 34–36]. The trust game with a reputation mechanism also approximates the situation of commerce conducted by the medieval Maghribi traders [37, 38] and the market for lemons (if a gossip-based reputation mechanism is operational) [1].

It was shown in a seminal paper that cooperation based on reputations is possible in the essentially asymmetric trust game [6]. Nevertheless, several important questions in this framework are theoretically unresolved. Is cooperation more ubiquitous than uncooperation, in particular when these two situations are both stable equilibria? How do different reputation-based strategies of buyers compete in a population? How does the reputation-based cooperation emerge through population dynamics? Non-evolutionary numerical results for the trust game with reputation mechanisms [39, 40] do not explain the stability and emergence of cooperation. An evolutionary theory [9, 25] and numerical simulations [41]

showed that reputations induce cooperation in the trust game. However, in these papers, each player serves the two roles with equal probability such that the game is essentially the symmetric prisoner's dilemma.

We theoretically clarify the possibility of reputation-based cooperation in the trust game by analyzing the Nash equilibria and the coevolutionary replicator dynamics of buyers and sellers. We show that coevolution of cooperative buyers and sellers is realized relatively easily. In particular, the fraction of buyers that score sellers does not have to be large, and many buyers do not discriminate between good and bad sellers even when cooperation prevails.

## Model

### Trust game

We analyze equilibria and evolutionary dynamics of the trust game (Fig. 1) in an infinite population of four types of buyers and two types of sellers under a reputation mechanism. Each seller possesses the binary reputation, good (G) or bad (B), that dynamically changes as a result of the single-shot trust game. Referring to the two reputations as G and B is purely conventional.

In a time unit, each seller  $i_s$  plays the trust game with a randomly chosen buyer  $i_b$ . The buyer  $i_b$  first decides whether to buy an item from  $i_s$  possibly on the basis of  $i_s$ 's reputation. If  $i_b$  does not buy, no transaction occurs, leaving the payoff 0 to both  $i_b$  and  $i_s$ . If  $i_b$  buys,  $i_s$  decides whether to ship the item (cooperate; C) or not (defect; D). If  $i_s$  ships the item, then both  $i_b$  and  $i_s$  obtain  $r$ . If  $i_s$  does not ship the item and exploits  $i_b$ ,  $i_s$  gains 1, and  $i_b$  gains  $-1$ . If  $r > 1$ , it is beneficial for both players to cooperate. If  $r < 0$ , transaction would never occur irrespective of the reputation mechanism. We set  $0 < r < 1$  to represent the social dilemma.

We repeat the procedure explained above for  $T$  time units such that each player plays the trust game  $T$  times on an average. We assume that  $1 \ll T \ll N$  such that the probability that the same pair of buyer and seller interact more than once within time  $T$  is infinitesimally small.

### Social norms

When  $i_b$  decides to buy,  $i_b$  assigns G or B to  $i_s$  on the basis of  $i_s$ 's action. The new reputation of  $i_s$  is instantaneously propagated to the entire population such that any buyer can refer to this information

when playing with  $i_s$  in later times. We refer to the rule according to which the buyer assigns a reputation to the seller as a social norm.

An intuitively rational social norm, which reputation mechanisms in successful online marketplaces apply, is to assign G and B when  $i_s$  has cooperated and defected, respectively. This norm is called image scoring [11, 12]. Alternatively,  $i_b$  may assign G no matter whether  $i_s$  cooperates or defects. We call this social norm indifferent scoring. A buyer is assumed to commit the assignment error with probability  $0 < \mu < 1/2$  such that the seller receives a reputation that is contrary to what is expected from the social norm.

In fact, some but not all of the buyers may be interested in rating sellers [2–4]. To investigate this scenario, we consider a case in which a buyer has a unique scoring type as well as strategy. We assume that a fraction of  $\theta$  and  $1 - \theta$  buyers are image scorers and indifferent scorers, respectively. We also assume that the scoring type does not affect the buyer’s payoff, such that the fraction of image scorers in the population does not change in the course of the replicator dynamics. Alternatively, we can assume that each buyer is a permanent indifferent scorer or permanent image scorer. This assumption may be controversial. We will return to this issue in Discussion.

We do not assume the reputation for buyers because in online marketplaces, the impact of the seller’s reputation is much larger than that of the buyer’s reputation [3, 5]. Previous laboratory experiments that modeled the situation in auction sites also neglected the reputation of buyers [7, 10, 34].

## Strategies

The probability that the buyer decides to buy from G and B sellers is denoted by  $b_G$  and  $b_B$ , respectively. We consider four strategies for buyers: unconditional buyer (Buy) specified by  $b_G = b_B = 1 - \epsilon$ , where  $0 < \epsilon < 1/2$ ; discriminator (Disc) specified by  $b_G = 1 - \epsilon$ ,  $b_B = \epsilon$ ; anti-discriminator (AntiDisc) specified by  $b_G = \epsilon$ ,  $b_B = 1 - \epsilon$ ; and unconditional no-buyer (NoBuy) specified by  $b_G = b_B = \epsilon$ .

As introduced in section “Trust game”, sellers have two strategies, C and D. For simplicity, we assume that the seller’s action is deterministic, whereas the buyer’s action is stochastic.

## Results

### Payoffs

The reputation of a seller obeys a Markov chain with two states G and B. To obtain the payoffs, we adopt the deterministic calculations developed by Ohtsuki and Iwasa [15, 18, 19].

If  $t$  ( $0 \leq t \leq T$ ) is sufficiently large, we can approximate the probability of the G reputation for a C seller and that for a D seller by the population averages denoted by  $\rho_C$  and  $\rho_D$ , respectively. The dynamics of the reputation averaged over the C sellers and that averaged over the D sellers are represented as

$$\frac{d\rho_C}{dt} = -\rho_C \left[ \theta \bar{b}_G^{\text{IM}} + (1 - \theta) \bar{b}_G^{\text{IN}} \right] \mu + (1 - \rho_C) \left[ \theta \bar{b}_B^{\text{IM}} + (1 - \theta) \bar{b}_B^{\text{IN}} \right] (1 - \mu), \quad (1)$$

$$\frac{d\rho_D}{dt} = -\rho_D \left[ \theta \bar{b}_G^{\text{IM}} (1 - \mu) + (1 - \theta) \bar{b}_G^{\text{IN}} \mu \right] + (1 - \rho_D) \left[ \theta \bar{b}_B^{\text{IM}} \mu + (1 - \theta) \bar{b}_B^{\text{IN}} (1 - \mu) \right], \quad (2)$$

where  $\bar{b}_G^a$  and  $\bar{b}_B^a$  ( $a = \text{IN}$  for indifferent scorer or  $\text{IM}$  for image scorer) are the probabilities that the buyer of scoring type  $a$  decides to buy from a G and B seller, respectively. For example,  $\left[ \theta \bar{b}_G^{\text{IM}} + (1 - \theta) \bar{b}_G^{\text{IN}} \right]$  on the right-hand side of Eq. (1) represents the probability that a buyer decides to buy from a C seller. Because the multiplicative factor  $\mu$  represents the probability that the C seller that has cooperated mistakenly receives B reputation, the first term on the right-hand side of Eq. (1) represents the case where the reputation of C seller turns from G to B. Similarly, the second term represents the case where the reputation of C seller turns from B to G. It should be noted that the seller's reputation does not change when the buyer decides not to buy.

The probabilities that the buyer decides to buy from a G and B seller are respectively represented as

$$\bar{b}_G^a = (1 - \epsilon)(y_1^a + y_2^a) + \epsilon(y_3^a + y_4^a), \quad (3)$$

$$\bar{b}_B^a = (1 - \epsilon)(y_1^a + y_3^a) + \epsilon(y_2^a + y_4^a), \quad (4)$$

where  $y_i^{\text{IN}}$  and  $y_i^{\text{IM}}$  are the fractions of buyers with strategy  $i$  among the indifferent scorers and among the image scorers, respectively. Buy, Disc, AntiDisc, and NoBuy correspond to  $i = 1, 2, 3$ , and 4, respectively;  $\sum_{i=1}^4 y_i^{\text{IN}} = \sum_{i=1}^4 y_i^{\text{IM}} = 1$ .

For simplicity, we assume that

$$y_i \equiv y_i^{\text{IN}} = y_i^{\text{IM}} \quad (1 \leq i \leq 4) \quad (5)$$

is initially satisfied. In other words, the scoring type and strategy are independent of each other. Note that  $0 \leq y_i \leq 1$  ( $1 \leq i \leq 4$ ) and  $\sum_{i=1}^4 y_i = 1$ . Because Eqs. (3), (4), and (5) imply  $\bar{b}_G \equiv \bar{b}_G^{\text{IN}} = \bar{b}_G^{\text{IM}}$  and  $\bar{b}_B \equiv \bar{b}_B^{\text{IN}} = \bar{b}_B^{\text{IM}}$ , Eqs. (1) and (2) give the limit values

$$\rho_C^* = \frac{(1 - \mu)\bar{b}_B}{\mu\bar{b}_G + (1 - \mu)\bar{b}_B}, \quad (6)$$

$$\rho_D^* = \frac{c_2\bar{b}_B}{c_1\bar{b}_G + c_2\bar{b}_B}, \quad (7)$$

where

$$\bar{b}_G = (1 - \epsilon)(y_1 + y_2) + \epsilon(y_3 + y_4), \quad (8)$$

$$\bar{b}_B = (1 - \epsilon)(y_1 + y_3) + \epsilon(y_2 + y_4), \quad (9)$$

and we set

$$c_1 = \theta(1 - \mu) + (1 - \theta)\mu, \quad (10)$$

$$c_2 = \theta\mu + (1 - \theta)(1 - \mu), \quad (11)$$

for notational convenience. Equation (6) does not depend on  $\theta$ , which reflects the fact that both the indifferent and image scorers evaluate C sellers as G with a large probability  $1 - \mu$  ( $> 1/2$ ). In contrast, Eq. (7) implies that  $\rho_D^*$  decreases with  $\theta$ . This is because the image scorer may issue a B reputation with a large probability, whereas the indifferent scorer does not.

For sufficiently large  $T$ , the reputations are in the equilibrium almost all the time. Then, the buyer's payoff should be equal to

$$\begin{aligned} P_i^b(b_G, b_B) &= rx [\rho_C^* b_G + (1 - \rho_C^*) b_B] - (1 - x) [\rho_D^* b_G + (1 - \rho_D^*) b_B] \\ &= \frac{rx [(1 - \mu)b_G \bar{b}_B + \mu \bar{b}_G b_B]}{\mu \bar{b}_G + (1 - \mu) \bar{b}_B} - \frac{(1 - x)(c_2 b_G \bar{b}_B + c_1 \bar{b}_G b_B)}{c_1 \bar{b}_G + c_2 \bar{b}_B}, \end{aligned} \quad (12)$$

where

$$(b_G, b_B) = \begin{cases} (1 - \epsilon, 1 - \epsilon), & \text{if } i = 1 \text{ (Buy)}, \\ (1 - \epsilon, \epsilon), & \text{if } i = 2 \text{ (Disc)}, \\ (\epsilon, 1 - \epsilon), & \text{if } i = 3 \text{ (AntiDisc)}, \\ (\epsilon, \epsilon), & \text{if } i = 4 \text{ (NoBuy)}, \end{cases} \quad (13)$$

and  $x$  ( $0 \leq x \leq 1$ ) is the fraction of C sellers. The payoffs for C and D sellers are given by

$$P_C^s = r [\rho_C^* \bar{b}_G + (1 - \rho_C^*) \bar{b}_B] = \frac{r \bar{b}_G \bar{b}_B}{\mu \bar{b}_G + (1 - \mu) \bar{b}_B} \quad (14)$$

and

$$P_D^s = \rho_D^* \bar{b}_G + (1 - \rho_D^*) \bar{b}_B = \frac{\bar{b}_G \bar{b}_B}{c_1 \bar{b}_G + c_2 \bar{b}_B}, \quad (15)$$

respectively.

Even if we consider the stochastic dynamics of the reputation and the buyer's action, the values of the payoffs derived above give the precise mean values.

**Proposition 1:** Regardless of the initial reputation value of a seller, in the limit  $T \rightarrow \infty$  and  $T/N \rightarrow 0$ , the expected payoff for the buyer is given by Eq. (12) and that for the seller is given by Eqs. (14) and (15).

## Proof of Proposition 1

The reputation score of the seller obeys a Markov chain with two states. Consider a C seller  $i_s$  with reputation G. The G reputation does not change in one time step if the paired buyer  $i_b$  decides to buy with probability  $\bar{b}_G$  and correctly assign G to  $i_s$  with probability  $1 - \mu$  or,  $i_b$  decides not to buy with probability  $1 - \bar{b}_G$ . Otherwise,  $i_s$ 's reputation turns into B. When  $i_s$  has reputation B, it is unchanged if  $i_b$  decides to buy with probability  $\bar{b}_B$  and commit assignment error with probability  $\mu$ , or  $i_b$  decides not to buy with probability  $1 - \bar{b}_B$ . Otherwise,  $i_s$ 's reputation turns into G.



Therefore, the transition matrix of the Markov chain is represented as

$$M \equiv \begin{pmatrix} (1-\mu)\bar{b}_G + (1-\bar{b}_G) & \mu\bar{b}_G \\ (1-\mu)\bar{b}_B & \mu\bar{b}_B + (1-\bar{b}_B) \end{pmatrix}, \quad (16)$$

where  $M_{ij}$  ( $1 \leq i, j \leq 2$ ) represents the transition probability from state  $i$  to state  $j$ , and we associate G and B with states 1 and 2, respectively. Because  $M$  is a nondegenerate (right) stochastic matrix, we can decompose  $M$  using the left and right eigenvectors corresponding to eigenvalue 1 as

$$M = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} \frac{(1-\mu)\bar{b}_B}{\mu\bar{b}_G + (1-\mu)\bar{b}_B} & \frac{\mu\bar{b}_G}{\mu\bar{b}_G + (1-\mu)\bar{b}_B} \end{pmatrix} + \lambda \mathbf{u} \mathbf{v}, \quad (17)$$

where  $-1 < \lambda < 1$  is the other eigenvalue of  $M$ , and  $\mathbf{v}$  and  $\mathbf{u}$  are the left and right eigenvectors corresponding to  $\lambda$ , respectively. We do not calculate  $\lambda$ ,  $\mathbf{v}$ , and  $\mathbf{u}$  because their values are immaterial in the following arguments. Note that the eigenvectors are normalized such that

$$\begin{pmatrix} \frac{(1-\mu)\bar{b}_B}{\mu\bar{b}_G + (1-\mu)\bar{b}_B} & \frac{\mu\bar{b}_G}{\mu\bar{b}_G + (1-\mu)\bar{b}_B} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \mathbf{v} \mathbf{u} = 1, \quad (18)$$

$$\mathbf{v} \begin{pmatrix} 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{(1-\mu)\bar{b}_B}{\mu\bar{b}_G + (1-\mu)\bar{b}_B} & \frac{\mu\bar{b}_G}{\mu\bar{b}_G + (1-\mu)\bar{b}_B} \end{pmatrix} \mathbf{u} = 0. \quad (19)$$

Assume that the C seller  $i_s$  initially has reputation G and B with probability  $p_{G,\text{init}}$  and  $p_{B,\text{init}}$ , respectively, where  $p_{G,\text{init}} + p_{B,\text{init}} = 1$ . Then, the probability that  $i_s$ 's reputation is G and B after playing  $t$  games is given by the first and second columns of  $(p_{G,\text{init}} \ p_{B,\text{init}})M^t$ , respectively.

The expected payoff for  $i_s$  in a single game is equal to  $r$  multiplied by the probability that  $i_b$  decides to buy. Therefore, the expected payoff for  $i_s$  per single game, averaged over  $1 \leq t \leq T$ , converges in the

limit  $T \rightarrow \infty$  to

$$\lim_{T \rightarrow \infty} \frac{1}{T} \begin{pmatrix} p_{G,\text{init}} & p_{B,\text{init}} \end{pmatrix} (I + M + M^2 + \dots + M^{T-1}) \begin{pmatrix} r\bar{b}_G \\ r\bar{b}_B \end{pmatrix} \quad (20)$$

$$= \lim_{T \rightarrow \infty} \frac{1}{T} \begin{pmatrix} p_{G,\text{init}} & p_{B,\text{init}} \end{pmatrix} \left[ T \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} \frac{(1-\mu)\bar{b}_B}{\mu\bar{b}_G + (1-\mu)\bar{b}_B} & \frac{\mu\bar{b}_G}{\mu\bar{b}_G + (1-\mu)\bar{b}_B} \end{pmatrix} + \frac{1-\lambda^T}{1-\lambda} \mathbf{uv} \right] \begin{pmatrix} r\bar{b}_G \\ r\bar{b}_B \end{pmatrix} \quad (21)$$

$$= \begin{pmatrix} p_{G,\text{init}} & p_{B,\text{init}} \end{pmatrix} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \begin{pmatrix} \frac{(1-\mu)\bar{b}_B}{\mu\bar{b}_G + (1-\mu)\bar{b}_B} & \frac{\mu\bar{b}_G}{\mu\bar{b}_G + (1-\mu)\bar{b}_B} \end{pmatrix} \begin{pmatrix} r\bar{b}_G \\ r\bar{b}_B \end{pmatrix} = \frac{r\bar{b}_G\bar{b}_B}{\mu\bar{b}_G + (1-\mu)\bar{b}_B}, \quad (22)$$

which reproduces Eq. (14). The expected payoff for the D seller, given by Eq. (15), can be derived in the same manner.

Next, we calculate the payoff for a Buy buyer  $i_b$ . In each time step, the expected number of C seller with reputation G and B with whom  $i_b$  is paired is equal to the first and second columns of  $x(p_{G,\text{init}} \ p_{B,\text{init}})M^t$ , respectively. When the C seller  $i_s$  has reputation G (B), the expected payoff for  $i_b$  in a single game is equal to  $r\bar{b}_G$  ( $r\bar{b}_B$ ). Therefore, the contribution of the C seller to the expected payoff for  $i_b$  per single game, averaged over  $T$  games, converges to

$$\lim_{T \rightarrow \infty} \frac{1}{T} x \begin{pmatrix} p_{G,\text{init}} & p_{B,\text{init}} \end{pmatrix} (I + M + M^2 + \dots + M^{T-1}) \begin{pmatrix} r\bar{b}_G \\ r\bar{b}_B \end{pmatrix} = \frac{xr\bar{b}_G\bar{b}_B}{\mu\bar{b}_G + (1-\mu)\bar{b}_B}. \quad (23)$$

This quantity is equal to the first term on the right-hand side of Eq. (12) when  $b_G = b_B = 1 - \epsilon$  (i.e., Buy). Analogous calculations for the case of D seller yields the second term on the right-hand side of Eq. (12) when  $b_G = b_B = 1 - \epsilon$ . The payoff for Disc, AntiDisc, and NoBuy can be calculated in a similar manner. Here is the end of the proof.

## Nash equilibria

Based on the expected payoff determined by Proposition 1, we identify the equilibria of the game. In the analysis, we exploit the fact that  $P_i^b(b_G, b_B)$  is linear in  $b_G$  and  $b_B$ . There are three types of Nash equilibria. The so-called uncooperative equilibrium is composed of NoBuy and the D seller. In the so-called cooperative equilibrium, Buy and Disc are mixed in the buyer's strategy and the probability of C is large in the seller's strategy. The cooperative equilibrium corresponds to the situation in which buyers

and sellers do not repeatedly interact but trust each other on the basis of the reputation mechanism. When it exists, it coexists with the uncooperative equilibrium. The other equilibrium appears only for a singular parameter set. Therefore, we are not concerned with it in the later analysis.

**Proposition 2:** The asymmetric trust game with a reputation mechanism whose expected payoffs are defined by Eqs. (12), (14), and (15) possesses the following three types of Nash equilibria.

1. Uncooperative equilibrium: combination of NoBuy and  $x^* = 0$ . This pure-strategy Nash equilibrium is also strict.
2. Cooperative equilibrium: the mixture of Buy and Disc, with the probability of Buy and Disc being

$$y_1^* = \frac{(-\mu + rc_1)(1 - \epsilon) + (1 - \mu - rc_2)\epsilon}{(1 - \mu - rc_2)(1 - 2\epsilon)} \quad (24)$$

and  $y_2^* = 1 - y_1^*$ , respectively. The probability of C seller is given by

$$x^* = \frac{c_1}{c_1 + \mu}. \quad (25)$$

The cooperative equilibrium exists when

$$\frac{\mu(1 - \epsilon) + (1 - \mu)\epsilon}{c_1(1 - \epsilon) + c_2\epsilon} < r < 1. \quad (26)$$

The cooperative equilibrium is also asymptotically stable under the replicator dynamics. For completeness, the replicator dynamics of buyers and sellers are given by

$$\frac{dy_i}{dt} = y_i \left[ P_i^b(b_G, b_B) - \overline{P^b} \right], \quad (27)$$

$$\begin{aligned} \frac{dx}{dt} &= x (P_C^s - P_D^s) \\ &= x(1 - x) \bar{b}_G \bar{b}_B \left[ \frac{r}{\mu \bar{b}_G + (1 - \mu) \bar{b}_B} - \frac{1}{c_1 \bar{b}_G + c_2 \bar{b}_B} \right], \end{aligned} \quad (28)$$

respectively, where the buyer's mean payoff is given by

$$\overline{P^b} = \left[ \frac{rx}{\mu \bar{b}_G + (1 - \mu) \bar{b}_B} - \frac{1 - x}{c_1 \bar{b}_G + c_2 \bar{b}_B} \right] \bar{b}_G \bar{b}_B. \quad (29)$$

3. A singular equilibrium: combination of Disc buyer and a mixed seller's strategy with any  $x$  satisfying

$$\frac{c_2}{1 - \mu + c_2} \leq x \leq \frac{c_1}{\mu + c_1}. \quad (30)$$

This equilibrium exists when

$$\frac{r}{\mu(1 - \epsilon) + (1 - \mu)\epsilon} - \frac{1}{c_1(1 - \epsilon) + c_2\epsilon} = 0. \quad (31)$$

We remark that the cooperative equilibrium is called so because  $\lim_{\mu \rightarrow 0} x^* = 1$ . We prove Proposition 2 in the next section.

It should be noted that extending the concept of the evolutionary stability to the asymmetric game is not straightforward. In the matrix game, a strictly (i.e., completely) mixed Nash equilibrium cannot be an asymptotically stable equilibrium under the replicator dynamics, and an evolutionarily stable strategy in the asymmetric game is necessarily a (pure) strict Nash equilibrium [42–45]. Nevertheless, Proposition 2 dictates that the cooperative equilibrium is an asymptotically stable strictly mixed strategy. This is possible because the payoff values are density-dependent in our model; it is not a matrix game. Therefore, we directly prove that the cooperative equilibrium is asymptotically stable in the replicator dynamics.

## Proof of Proposition 2

We identify all the mixed-strategy Nash equilibria of the asymmetric game whose payoffs are given by Eqs. (12), (14), and (15).

### One buyer's strategy

Consider a possible equilibrium composed of a single buyer's strategy. If there is only Buy, AntiDisc, or NoBuy, we substitute  $(\bar{b}_G, \bar{b}_B) = (1 - \epsilon, 1 - \epsilon)$ ,  $(\epsilon, 1 - \epsilon)$ , and  $(\epsilon, \epsilon)$ , respectively, in Eqs. (14) and (15) to obtain  $P_C^s - P_D^s < 0$  for  $r < 1$ . Therefore,  $x^* = 0$  must be satisfied in a possible Nash equilibrium.

When  $x^* = 0$ , Eq. (12) is simplified to

$$P_i^b(b_G, b_B) = -\frac{c_2 b_G \bar{b}_B + c_1 \bar{b}_G b_B}{c_1 \bar{b}_G + c_2 \bar{b}_B}, \quad (32)$$

where  $c_1$  and  $c_2$  are defined in Eqs. (10) and (11), respectively. Equation (32) implies that the payoff for

NoBuy is larger than those for Buy, Disc, and AntiDisc. Therefore, the combination of NoBuy and D seller is the only (strict) Nash equilibrium allowed in this regime. We call this equilibrium the uncooperative equilibrium.

Suppose instead that there is only Disc. If

$$\frac{r}{\mu \bar{b}_G + (1 - \mu) \bar{b}_B} - \frac{1}{c_1 \bar{b}_G + c_2 \bar{b}_B} = 0, \quad (33)$$

where  $\bar{b}_G = 1 - \epsilon$  and  $\bar{b}_B = \epsilon$ , we obtain  $P_C^s = P_D^s$  for any  $x$ . Substituting Eq. (33) in Eq. (12) yields

$$P_i^b(b_G, b_B) = \frac{[x(1 - \mu) - (1 - x)c_2] b_G \bar{b}_B + [x\mu - (1 - x)c_1] \bar{b}_G b_B}{c_1 \bar{b}_G + c_2 \bar{b}_B}. \quad (34)$$

Equation (34) indicates that Disc obtains a payoff larger than or equal to those of Buy, AntiDisc, and NoBuy if

$$\frac{c_2}{1 - \mu} \leq \frac{x}{1 - x} \leq \frac{c_1}{\mu}. \quad (35)$$

The range of  $x$  that satisfies Eq. (35) always exists because  $\mu < 1/2$  guarantees  $c_2/(1 - \mu) < c_1/\mu$ . Therefore, the combination of Disc and Eq. (35), i.e.,

$$\frac{c_2}{1 - \mu + c_2} \leq x \leq \frac{c_1}{\mu + c_1}, \quad (36)$$

yields Nash equilibria.

If Eq. (33) is not satisfied, which is a generic case, only  $x = 0$  and  $x = 1$  may result in a Nash equilibrium because the difference between Eqs. (14) and (15) is independent of  $x$ . If  $x^* = 0$  (i.e.,  $P_C^s < P_D^s$ ), Eq. (12) is simplified to

$$P_i^b(b_G, b_B) = -\frac{c_2 \epsilon b_G + c_1 (1 - \epsilon) b_B}{c_1 (1 - \epsilon) + c_2 \epsilon}. \quad (37)$$

Equation (37) implies that NoBuy gains a larger payoff than Disc. If  $x^* = 1$ , substituting  $\bar{b}_G = 1 - \epsilon$  and  $\bar{b}_B = \epsilon$  in Eqs. (14) and (15) and setting  $P_C^s > P_D^s$  lead to the following necessary condition for Disc to be Nash:

$$r > \frac{\mu(1 - \epsilon) + (1 - \mu)\epsilon}{c_1(1 - \epsilon) + c_2\epsilon}. \quad (38)$$

Equation (38) replicates Eq. (26). When Eq. (38) is satisfied, substituting  $x^* = 1$ ,  $\bar{b}_G = 1 - \epsilon$ , and  $\bar{b}_B = \epsilon$  in Eq. (12) yields

$$P_i^b(b_G, b_B) = \frac{r[(1-\mu)\epsilon b_G + \mu(1-\epsilon)b_B]}{\mu(1-\epsilon) + (1-\mu)\epsilon}. \quad (39)$$

Therefore, Buy gains a larger payoff than Disc, such that the pure Disc is not Nash.

In conclusion, the only pure (strict) Nash equilibrium is the uncooperative equilibrium composed of NoBuy and D seller.

### Mixture of two buyer's strategies

Consider the mixed strategies (i.e., coexistence) of two buyer's strategies as candidates of Nash equilibria. If we select two strategies out of Buy, AntiDisc, and NoBuy, we can show  $P_C^s < P_D^s$  in a manner similar to the case of the one buyer's strategy. In this case,  $x^* = 0$  must hold true in the equilibrium. Equation (12) with  $x^* = 0$  indicates that the payoff for NoBuy is larger than that for AntiDisc, which is larger than that for Buy. Therefore, such a mixed-strategy Nash equilibrium, in which the payoff for the two strategies of buyers must be the same, does not exist. This implies that Disc must be selected as one of the two buyer's strategies in a possible Nash equilibrium.

*Mixture of Buy and Disc: cooperative equilibrium:* Consider a mixture of Buy and Disc, which we call the cooperative equilibrium. Note that

$$\bar{b}_G = 1 - \epsilon, \quad (40)$$

$$\bar{b}_B = (1 - \epsilon)y_1 + \epsilon y_2 = (1 - 2\epsilon)y_1 + \epsilon. \quad (41)$$

Because  $P_i^b(1 - \epsilon, 1 - \epsilon) = P_i^b(1 - \epsilon, \epsilon)$  must be satisfied in the Nash equilibrium, the coefficient of  $b_B$  in Eq. (12) must be equal to 0. This condition combined with  $P_C^s = P_D^s$  yields

$$\bar{b}_B^* = \frac{(-\mu + rc_1)(1 - \epsilon)}{1 - \mu - rc_2} \quad (42)$$

and

$$x^* = \frac{c_1}{c_1 + \mu}. \quad (43)$$

Equation (43) replicates Eq. (25).

The mixture of Buy and Disc is equivalent to  $\epsilon < \bar{b}_B^* < 1 - \epsilon$ . The inequality  $\bar{b}_B^* < 1 - \epsilon$  is always satisfied if  $r < 1$  (note that the denominator of Eq. (42) is always positive if  $\mu < 1/2$ ). The inequality  $\epsilon < \bar{b}_B^*$  is equivalent to Eq. (38). Given Eq. (38), the equilibrium probability of Buy (i.e.,  $y_1^*$ ) and that of Disc (i.e.,  $y_2^* = 1 - y_1^*$ ) in the cooperative equilibrium, shown in Eq. (24), are derived by substituting Eq. (42) in

$$\bar{b}_B^* = (1 - \epsilon)y_1^* + \epsilon y_2^* = (1 - 2\epsilon)y_1^* + \epsilon. \quad (44)$$

To show the stability of the cooperative equilibrium, we first compare  $P_i^b(1 - \epsilon, 1 - \epsilon) = P_i^b(1 - \epsilon, \epsilon)$ , i.e., the payoff for Buy and that for Disc, and  $P_i^b(\epsilon, \epsilon)$ , i.e., the payoff for NoBuy, in the cooperative equilibrium. NoBuy gains a smaller payoff than Buy and Disc if the coefficient of  $b_G$  in Eq. (12) is positive in the cooperative equilibrium. The substitution of Eqs. (42) and (43) in Eq. (12) suggests that this condition is equivalent to

$$1 - \mu > \frac{\mu c_2}{c_1}. \quad (45)$$

Equation (45) is equivalent to  $\mu < 1/2$ , which we have assumed. Therefore, NoBuy gains a smaller payoff than Buy and Disc. Because  $P_i^b(1 - \epsilon, 1 - \epsilon) = P_i^b(1 - \epsilon, \epsilon)$  implies  $P_i^b(\epsilon, 1 - \epsilon) = P_i^b(\epsilon, \epsilon)$ , AntiDisc also gains a smaller payoff than Buy and Disc in the cooperative equilibrium.

Consider the invariant subspace of the strategy space where only Buy and Disc buyers (and C and D sellers) exist. The mixed strategy specified by  $(y_1, x) = (y_1^*, x^*)$  is a Nash equilibrium when the buyer's strategies are restricted to either Buy and Disc because  $P_i^b(1 - \epsilon, 1 - \epsilon) = P_i^b(1 - \epsilon, \epsilon)$  and  $P_C^s = P_D^s$ . Because  $P_i^b(\epsilon, 1 - \epsilon) = P_i^b(\epsilon, \epsilon) < P_i^b(1 - \epsilon, 1 - \epsilon) = P_i^b(1 - \epsilon, \epsilon)$ ,  $(y_1, x) = (y_1^*, x^*)$  is a Nash equilibrium when all the four types of buyer's strategies are allowed.

The inequality  $P_i^b(\epsilon, 1 - \epsilon) = P_i^b(\epsilon, \epsilon) < P_i^b(1 - \epsilon, 1 - \epsilon) = P_i^b(1 - \epsilon, \epsilon)$  also assures that, under the replicator dynamics, the cooperative equilibrium is asymptotically stable against the introduction of an infinitesimal fraction of AntiDisc or NoBuy. Therefore, we are left to show the asymptotic stability of the cooperative equilibrium within the abovementioned two-dimensional subspace parametrized by  $y_1$  and  $x$ . For the sake of the linear stability analysis, we take  $\bar{b}_B$  and  $x$  as the independent variables and linearize Eqs. (27) with  $i = 1$  and (28) and substitute the one-to-one relationship between  $y_1$  and  $\bar{b}_B$  given by

Eq. (41) in Eq. (27). The Jacobian in the equilibrium is given by

$$J = \begin{pmatrix} \frac{\partial}{\partial \bar{b}_B} \frac{d\bar{b}_B}{dt} & \frac{\partial}{\partial x} \frac{d\bar{b}_B}{dt} \\ \frac{\partial}{\partial \bar{b}_B} \frac{dx}{dt} & \frac{\partial}{\partial x} \frac{dx}{dt} \end{pmatrix}, \quad (46)$$

where all the derivatives are evaluated at  $(\bar{b}_B, x) = (\bar{b}_B^*, x^*)$ , and

$$\frac{\partial}{\partial \bar{b}_B} \frac{d\bar{b}_B}{dt} = (\bar{b}_B^* - \epsilon)(1 - \epsilon)(1 - \epsilon - \bar{b}_B^*) \left\{ \frac{-rx^*\mu(1 - \mu)}{[\mu(1 - \epsilon) + (1 - \mu)\bar{b}_B^*]^2} + \frac{(1 - x^*)c_1c_2}{[c_1(1 - \epsilon) + c_2\bar{b}_B^*]^2} \right\}, \quad (47)$$

$$\frac{\partial}{\partial x} \frac{d\bar{b}_B}{dt} = (\bar{b}_B^* - \epsilon)(1 - \epsilon)(1 - \epsilon - \bar{b}_B^*) \left\{ \frac{r\mu}{\mu(1 - \epsilon) + (1 - \mu)\bar{b}_B^*} + \frac{c_1}{c_1(1 - \epsilon) + c_2\bar{b}_B^*} \right\}, \quad (48)$$

$$\frac{\partial}{\partial \bar{b}_B} \frac{dx}{dt} = x^*(1 - x^*)(1 - \epsilon)\bar{b}_B^* \left\{ \frac{-r(1 - \mu)}{[\mu(1 - \epsilon) + (1 - \mu)\bar{b}_B^*]^2} + \frac{c_2}{[c_1(1 - \epsilon) + c_2\bar{b}_B^*]^2} \right\}, \quad (49)$$

$$\frac{\partial}{\partial x} \frac{dx}{dt} = 0, \quad (50)$$

The necessary and sufficient condition for the cooperative equilibrium to be stable under the replicator dynamics is given by  $\text{trace} J < 0$  and  $\det J > 0$ . By substituting Eqs. (42) and (43) in Eq. (47), we obtain

$$\text{trace} J = \frac{\mu(\bar{b}_B^* - \epsilon)(1 - \epsilon - \bar{b}_B^*)c_1(1 - \mu - rc_2)^2}{(1 - \epsilon)(c_1 + \mu)(c_1 - \mu)^2} \left( -\frac{1 - \mu}{r} + c_2 \right). \quad (51)$$

Because the right-hand side of Eq. (49) is positive, we obtain

$$\det J \propto -\frac{\partial}{\partial \bar{b}_B} \frac{dx}{dt} = \frac{x^*(1 - x^*)\bar{b}_B^*(1 - \mu - rc_2)^2}{(1 - \epsilon)(c_1 - \mu)^2} \left( \frac{1 - \mu}{r} - c_2 \right). \quad (52)$$

Equations (51) and (52) suggest that the cooperative equilibrium is stable if and only if

$$\frac{1 - \mu}{r} > c_2. \quad (53)$$

Equation (53) is satisfied for any  $r$  ( $0 < r < 1$ ) if  $\mu < 1/2$ . Therefore, the cooperative equilibrium is asymptotically stable under the replicator dynamics.

*Mixture of Disc and AntiDisc:* If Disc and AntiDisc are mixed in the equilibrium,

$$P_i^b(1 - \epsilon, \epsilon) = P_i^b(\epsilon, 1 - \epsilon) \quad (54)$$



holds true. Because  $P_i^b(b_G, b_B)$  is linear in  $b_G$  and  $b_B$ , the coefficient of  $b_G$  and that of  $b_B$  must be the same for Eq. (54) to be satisfied. If both coefficients are positive, the payoff for Buy is larger than that for Disc and AntiDisc in this equilibrium. If both coefficients are negative, the payoff for NoBuy is larger than that for Disc and AntiDisc in the equilibrium. In either case, the mixture of Disc and AntiDisc cannot be Nash.

*Mixture of Disc and NoBuy:* If Disc and NoBuy are mixed in the equilibrium, we obtain  $P_i^b(1 - \epsilon, \epsilon) = P_i^b(\epsilon, \epsilon)$  and  $\bar{b}_B = b_B = \epsilon$ . Therefore, the coefficient of  $b_G$  in Eq. (12), which we denote by  $h(y_2)$  as a function of the density of Disc, is represented as

$$h(y_2) = \frac{rx(1 - \mu)\epsilon}{\mu\bar{b}_G + (1 - \mu)\epsilon} - \frac{(1 - x)c_2\epsilon}{c_1\bar{b}_G + c_2\epsilon}, \quad (55)$$

where

$$\bar{b}_G = (1 - \epsilon)y_2 + \epsilon(1 - y_2) \quad (56)$$

must be equal to 0 in the equilibrium. From  $h(y_2) = 0$  and  $P_C^s = P_D^s$ , we obtain

$$\bar{b}_G^* = \frac{1 - \mu - rc_2}{-\mu + rc_1} \quad (57)$$

and

$$x^* = \frac{c_2}{1 - \mu + c_2}. \quad (58)$$

If  $dh(y_2)/dy_2 > 0$  in the equilibrium, Disc (NoBuy) in a mixed strategy in which there are slightly more (less) probability of Disc than in the equilibrium  $(\bar{b}_G, x) = (\bar{b}_G^*, x^*)$  obtains a larger payoff than NoBuy (Disc). In this case,  $(\bar{b}_G^*, x^*)$  is not Nash because such a slightly modified mixed strategy of the buyer obtains a larger payoff than  $(\bar{b}_G^*, x^*)$ . On the basis of the relationship  $dh/dy_2 = (1 - 2\epsilon)dh/d\bar{b}_G$ , which is derived from Eq. (56), we rewrite  $h(y_2)$  as  $h(\bar{b}_G)$  and examine  $dh(\bar{b}_G)/d\bar{b}_G$  in the equilibrium (note that we assumed  $\epsilon < 1/2$ ). Using the fact that the right-hand side of Eq. (55) is equal to 0 when

$(\bar{b}_G, x) = (\bar{b}_G^*, x^*)$ , we obtain

$$\begin{aligned} \left. \frac{dh(b_G)}{db_G} \right|_{(\bar{b}_G, x) = (\bar{b}_G^*, x^*)} &= \frac{-rx^*(1-\mu)\mu\epsilon}{[\mu\bar{b}_G^* + (1-\mu)\epsilon]^2} + \frac{(1-x^*)c_1c_2\epsilon}{(c_1\bar{b}_G^* + c_2\epsilon)^2} \\ &= \frac{rx^*(1-\mu)(1-2\mu)\epsilon^2\theta}{[\mu\bar{b}_G^* + (1-\mu)\epsilon]^2(c_1\bar{b}_G^* + c_2\epsilon)} \\ &> 0 \end{aligned} \quad (59)$$

if  $\mu < 1/2$ . Therefore, the mixture of Disc and NoBuy is not Nash.

### Mixture of three or four buyer's strategies

If three or four buyer's strategies are mixed in an equilibrium, their payoffs must be identical. Therefore, the coefficient of  $b_G$  and that of  $b_B$  in Eq. (12) must be equal to 0, which requires  $\theta(2\mu - 1) = 0$ . Because  $\mu < 1/2$ , this relationship is not satisfied when the image scorer exists (i.e.,  $\theta > 0$ ).

When  $\theta = 0$ , we obtain  $P_C^s < P_D^s$  by substituting  $c_1 = \mu$  and  $c_2 = 1 - \mu$  (derived from Eqs. (10) and (11)) in Eqs. (14) and (15). Therefore,  $x^* = 0$  must hold in the equilibrium. In this situation, Eq. (12) is reduced to

$$P_i^b(b_G, b_B) = -\frac{(1-\mu)b_G\bar{b}_B + \mu\bar{b}_Gb_B}{\mu\bar{b}_G + (1-\mu)\bar{b}_B}. \quad (60)$$

Equation (60) indicates that NoBuy's payoff is larger than Disc's and AntiDisc's payoffs, which are larger than Buy's payoff. Consequently, three or four buyers' strategies cannot be mixed in an equilibrium. Here is the end of the proof.

### Indifferent scoring

In a case with only indifferent scorers (i.e.,  $\theta = 0$ ), Eq. (26) is never satisfied (and Eq. (31) is not satisfied, either). Therefore, the uncooperative equilibrium is the only Nash equilibrium. This outcome is expected because, under indifferent scoring, the players perform the usual trust game [28–31].

### Image scoring

When there are only image scorers (i.e.,  $\theta = 1$ ), the cooperative equilibrium is realized in a wide parameter region because Eq. (26) with  $\theta = 1$  and  $\mu \rightarrow 0$  is reduced to  $\epsilon/(1-\epsilon) < r < 1$ ;  $\epsilon < 1/2$  is the probability

that the buyer misimplements the action. In the limit  $\mu \rightarrow 0$ , the equilibrium probability of Buy (i.e.,  $y_1^*$ ) is plotted as a function of  $\epsilon$  and  $r$  in Fig. 2(a). In the parameter region in which  $y_1^* = 0$  (black region in Fig. 2(a)), the cooperative equilibrium does not exist. In the cooperative equilibrium, the fraction of Buy is large for a large  $r$  or a small  $\epsilon$ . In particular,  $\lim_{\mu \rightarrow 0, r \rightarrow 1} y_1^* = 1$  irrespective of the value of  $\epsilon$ . In the limit  $r \rightarrow 1$ , the trust game is a weak social dilemma such that the D seller's payoff is only infinitesimally larger than the C seller's payoff. The advantage of the D seller is offset by the B reputation that the D seller receives from just a small fraction of Disc buyers (i.e.,  $y_2 \ll 1$ ). Even if a majority of buyers is nondiscriminative Buy, cooperation between buyers and sellers can be sustained by the reputation-regarding behavior of a small fraction of Disc.

For a general value of  $\theta$ , suppose that Eq. (26) is satisfied such that the cooperative equilibrium is Nash. For  $\epsilon = 0.1$  and  $\mu \rightarrow 0$ ,  $y_1^*$  in the cooperative equilibrium is shown for various values of  $\theta$  and  $r$  in Fig. 2(b).  $y_1^*$  increases as  $\theta$  and  $r$  increase.

For some values of  $\mu$  and  $\epsilon$ , the critical line obtained from Eq. (26) is plotted in Fig. 3. The cooperative equilibrium exists and is stable above the critical line. Figure 3 indicates that cooperation based on the reputation mechanism occurs in a large parameter region, particularly for large values of  $\theta$  and  $r$ . The critical line is rather insensitive to  $\mu$  and relatively sensitive to  $\epsilon$ . Nevertheless, cooperation is possible even for a large probability of the implementation error  $\epsilon = 0.2$ .

## Attractive basin of the cooperative and uncooperative equilibria

The cooperative and uncooperative equilibria can coexist. A fuller understanding of the model requires a global analysis of the replicator dynamics to determine which of the two equilibria is more likely to be attained. In addition, periodic or chaotic attractors may exist when a population evolves. To exclude this possibility and examine the attractive basin of the two equilibria, we numerically run the standard two-population replicator dynamics [43–45] from various initial conditions to identify the limit set of the dynamics. For fixed parameter values, we assume that the initial condition is distributed according to the uniform density on the state space, i.e.,  $\{(y_1, y_2, y_3, y_4, x) : y_i \geq 0 \ (1 \leq i \leq 4), \sum_{i=1}^4 y_i = 1, 0 \leq x \leq 1\}$ . It should be noted that Eq. (5) is preserved under the replicator dynamics because the buyer's payoff depends on the buyer's strategy but is independent of whether the buyer is an indifferent or image scorer. We have implicitly assumed in Eqs. (27) and (28) that sellers and buyers have identical adaptation rates. However, the following results are qualitatively the same even if the two adaptation rates are different.

The replicator dynamics are four-dimensional, with three degrees of freedom derived from the buyer's population and one degree of freedom derived from the seller's population. Note that the selection occurs separately among buyers and sellers.

For an expository purpose, we start with the system without AntiDisc. There are three strategies for buyers and two strategies for sellers. The replicator dynamics are three-dimensional, and the state space is a triangular prism defined by  $\{(y_1, y_2, y_4, x) : y_1, y_2, y_4 \geq 0, y_3 = 0, y_1 + y_2 + y_4 = 1, 0 \leq x \leq 1\}$ . For  $\mu = 0.02$ ,  $\epsilon = 0.1$ ,  $r = 0.15$ , and  $\theta = 1$ , initial conditions located below the boundary shown in Fig. 4 are attracted to the uncooperative equilibrium. We confirmed that all the complementary regions in the interior of the triangular prism are attracted to the cooperative equilibrium. Our numerical simulations strongly suggest that there is no other limit set.

To better quantify the possibility of cooperation, we measure the volume of the attractive basin of the cooperative equilibrium. The relative volume of the attractive basin in the triangular prism is shown in Fig. 5(a) for  $\mu = 0.02$ ,  $\epsilon = 0.1$ , and various values of  $r$  and  $\theta$ . The critical line for the existence and stability of the cooperative equilibrium, implied by Eq. (26), is shown by the solid line. We find that the cooperative equilibrium is attractive for a substantial variety of initial conditions as  $\theta$  and  $r$  increase.

In the presence of four strategies for buyers, the relative volume of the attractive basin of the cooperative equilibrium in the state space is shown in Fig. 5(b). The results are qualitatively the same as those shown in Fig. 5(a). We confirmed that the rest of the state space belongs to the attractive basin of the uncooperative equilibrium. The introduction of AntiDisc does not inhibit the evolution of cooperation.

## Discussion

For the trust game, we have shown that cooperation between unacquainted buyers and sellers can be established under the image scoring norm (i.e., reputation mechanism). In the cooperative equilibrium, the population of buyers (i.e., investors) is a mixture of Buy (unconditional buyer) and Disc (discriminator that decides whether to buy depending on the seller's reputation). The majority of sellers (i.e., trustees) reciprocates the buyer's trust although the sellers are not expected to meet the same buyers again. It should be noted that not every buyer discriminates between good and bad sellers in the cooperative equilibrium. This feature is shared by some previous models of indirect reciprocity [14, 17, 19]. The probability of discriminators can be small depending on the parameter values. In addition, the buyer's

reputation may be of little practical use for maintaining cooperative transactions. This claim is consistent with the previous theoretical result [6] and the empirical finding that reputation for a seller has a greater impact than that for a buyer [3, 5, 10].

Our model and results are distinct from previous ones obtained from the models using the symmetrized donation game, which we call indirect reciprocity games for now [8, 9, 11–16]. First, when cooperation prevails, a player with a good reputation is helped by unacquainted players in the indirect reciprocity games. In our model, a seller with a good reputation wins the trust, not explicit help, of unacquainted buyers. By reciprocating the buyer’s trust, the seller obtains a relatively large momentary payoff and a good reputation. Then, a good reputation elicits trust and long-term cooperation from buyers.

A second difference is in the consequence of the image scoring. In the indirect reciprocity games, defection against a bad player is regarded to be bad under the image scoring. Therefore, the image scoring does not yield cooperation [13–16]. In contrast, in our model, a discriminative buyer (i.e., Disc) that defects against (i.e., does not buy from) a bad seller does not receive a bad reputation. This is because buyers do not own reputation scores by definition; the asymmetry of roles in the trust game allows the image scoring to support cooperation. Although a previous paper discussed this issue before [46], we analytically derived the conditions under which cooperation based on the image scoring occurs. It should be noted that the image scoring norm is simpler than the social norms required for cooperation in the indirect reciprocity games [13–16, 18, 19].

In other models, the mere reputation mechanism enables cooperation among unacquainted players in the trust game [9, 25] and other games [9, 24, 26, 27]. However, the component social dilemma game in these studies is essentially symmetric. Our model is inherently asymmetric such that a player is either a permanent buyer or permanent seller. As compared to a recent paper in which impacts of asymmetric roles in the trust game are numerically studied [46], we analytically established the conditions under which reputation-based cooperation occurs in the asymmetric trust game. In online marketplaces [10], the market for lemons [1], and presumably many other transaction scenes, some people may participate entirely or mostly as buyers and others as sellers.

The constructive role of reputations for the trustee (i.e., seller) in asymmetric interactions was formulated in a classic paper many years ago [6]. Our contribution to the understanding of this established mechanism is that we have clarified competition among different strategies using evolutionary game theory.

We have investigated a scenario in which indifferent scorers, which do not essentially score sellers, and image scorers coexist in a buyer's population. Remarkably, the fraction of image scorers needed for cooperation is not large; Eq. (26) indicates that the threshold fraction of image scorers is equal to 0.60 for  $r = 0.2$  and 0.038 for  $r = 0.8$  when  $\mu = 0.02$  and  $\epsilon = 0.1$ . An alternative assumption for the behavior of indifferent scorers is that they do not alter sellers' scores rather than always give a good reputation to sellers. Analysis of this case is warranted for future work.

An important limitation of the present study is that the fraction of indifferent scorers and that of image scorers are invariant over time. In other words, indifferent and image scorers are assumed to receive the same payoff if the strategy (Buy, Disc, AntiDisc, or NoBuy) is the same. In fact, the image scoring may be more costly than the indifferent scoring because the image scorer has to know and report whether sellers cooperate or defect. Therefore, an image scorer may be tempted to turn into an indifferent scorer if the incentive to rate sellers is absent [2, 3, 10]. A similar cost is briefly mentioned in previous literature in the context of indirect reciprocity games [9, 47]. In our model, cooperation disappears if there are too many indifferent scorers. This result parallels with that for the indirect reciprocity games in which cooperation is not realized if there are too few observers in the one-shot game [11, 16]. In practice, rewarding image scorers and shutting down indifferent scorers' access to reputation information, for example, are means to circumvent the scoring cost [3]. We remark that the competition of social norms was analyzed in different models of indirect reciprocity [20–22].

In the context of asymmetric interaction between cleaner and client fishes, indirect reciprocity was investigated in an inherently asymmetric variant of the trust game with a binary internal state for the trustee [23]. Our model and results are distinct from theirs. In their model, a trustee reciprocates and attracts the investor when the trustee is in one particular state and exploits the investor by switching to the other state. The trustee does not steadily maintain a good reputation and uses the temporarily good reputation to exploit the investor. The authors acknowledge that their mechanism is different from the conventional concept of indirect reciprocity. In our model, the trustee (i.e., seller) steadily maintains a good reputation to invoke help from the investor (i.e., buyer). Our results suggest that, under appropriate conditions, the conventional indirect reciprocity may be established between cleaner and client fishes without resorting to the concept of the binary state. Finally, beyond the relevance to online transactions, our results provide a firm solution to the moral hazard problem that is represented by the trust game. Examples include offline markets [6] such as the market for lemons [1] and labor

markets [29,33]. Implementing a reputation mechanism only for the seller (i.e., investee) induces trust and cooperation between unacquainted individuals in the trust game. Good cooperative sellers and trustful buyers coevolve.

## Acknowledgments

We thank Michihiro Kandori for discussion and Hisashi Ohtsuki for providing valuable comments on the manuscript.

## References

1. Akerlof GA (1970) The market for “lemons”: Quality uncertainty and the market mechanism. *Quarterly J Econom* 84: 488–500.
2. Resnick P, Zeckhauser R, Friedman E, Kuwabara K (2000) Reputation systems. *Communications of the ACM* 43: 45–48.
3. Malaga RA (2001) Web-based reputation management systems: Problems and suggested solutions. *Electronic Commerce Res* 1: 403–417.
4. Dellarocas C (2003) The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Manage Sci* 49: 1407–1424.
5. Brown J, Morgan J (2006) Reputation in online auctions: the market for trust. *California Man Rev* 49: 61–81.
6. Klein B, Leffler KB (1981) The role of market forces in assuring contractual performance. *J Polit Economy* 89: 615–641.
7. Bolton GE, Katok E, Ockenfels A (2004) How effective are electronic reputation mechanisms? an experimental investigation. *Manage Sci* 50: 1587–1602.
8. Nowak MA, Sigmund K (2005) Evolution of indirect reciprocity. *Nature* 437: 1291–1298.
9. Sigmund K (2010) *The Calculus of Selfishness*. Princeton, NJ: Princeton University Press.

10. Resnick P, Zeckhauser R (2002) Trust among strangers in internet transactions: Empirical analysis of ebay's reputation system. *Adv Appl Microeconomics* 11: 127–157.
11. Nowak MA, Sigmund K (1998) Evolution of indirect reciprocity by image scoring. *Nature* 393: 573–577.
12. Nowak MA, Sigmund K (1998) The dynamics of indirect reciprocity. *J Theor Biol* 194: 561–574.
13. Leimar O, Hammerstein P (2001) Evolution of cooperation through indirect reciprocity. *Proc R Soc B* 268: 745–753.
14. Panchanathan K, Boyd R (2003) A tale of two defectors: the importance of standing for evolution of indirect reciprocity. *J Theor Biol* 224: 115–126.
15. Ohtsuki H, Iwasa Y (2004) How should we define goodness?—reputation dynamics in indirect reciprocity. *J Theor Biol* 231: 107–120.
16. Brandt H, Sigmund K (2004) The logic of reprobation: assessment and action rules for indirect reciprocation. *J Theor Biol* 231: 475–486.
17. Brandt H, Sigmund K (2006) The good, the bad and the discriminator — errors in direct and indirect reciprocity. *J Theor Biol* 239: 183–194.
18. Ohtsuki H, Iwasa Y (2006) The leading eight: social norms that can maintain cooperation by indirect reciprocity. *J Theor Biol* 239: 435–444.
19. Ohtsuki H, Iwasa Y (2007) Global analyses of evolutionary dynamics and exhaustive search for social norms that maintain cooperation by reputation. *J Theor Biol* 244: 518–531.
20. Chalub FACC, Santos FC, Pacheco JM (2006) The evolution of norms. *J Theor Biol* 241: 233–240.
21. Pacheco JM, Santos FC, Chalub FACC (2006) Stern-judging: A simple, successful norm which promotes cooperation under indirect reciprocity. *PLoS Comput Biol* 2: 1634–1638.
22. Uchida S, Sigmund K (2010) The competition of assessment rules for indirect reciprocity. *J Theor Biol* 263: 13–19.
23. Johnstone RA, Bshary R (2007) Indirect reciprocity in asymmetric interactions: When apparent altruism facilitates profitable exploitation. *Proc R Soc B* 274: 3175–3181.



24. Nowak MA, Page KM, Sigmund K (2000) Fairness versus reason in the ultimatum game. *Science* 289: 1773–1775.
25. Sigmund K, Hauert C, Nowak MA (2001) Reward and punishment. *Proc Natl Acad Sci USA* 98: 10757–10762.
26. Raub W, Weesie J (1990) Reputation and efficiency in social interactions: an example of network effects. *Am J Sociol* 96: 626–654.
27. Kandori M (1992) Social norms and community enforcement. *Rev of Econom Studies* 59: 63–80.
28. Dasgupta P (1988) Trust as a commodity. In: *Trust: making and breaking cooperative relations*, D Gambetta, ed : 49–72.
29. Kreps DM (1990) Cooperative culture and economic theory. In: *Perspectives on Positive Political Economy* (J Alt and K Shepsle, Eds) : 90–143.
30. Berg J, Dickhaut J, McCabe K (1995) Trust, reciprocity, and social history. *Games Econom Behav* 10: 122–142.
31. Fehr E, Fischbacher U (2003) The nature of human altruism. *Nature* 425: 785–791.
32. McCabe KA, Smith VL, LePore M (2000) Intentionality detection and “mindreading”: Why does game form matter? *Proc Natl Acad Sci USA* 97: 4404–4409.
33. McCabe KA, Rigdon ML, Smith VL (2003) Positive reciprocity and intentions in trust games. *J Econom Behav & Organization* 52: 267–275.
34. Keser C (2003) Experimental games for the design of reputation management systems. *IBM Syst J* 42: 498–506.
35. Basu S, Dickhaut J, Hecht G, Towry K, Waymire G (2009) Recordkeeping alters economic history by promoting reciprocity. *Proc Natl Acad Sci USA* 106: 1009–1014.
36. Bracht J, Feltovich N (2009) Whatever you say, your reputation precedes you: Observation and cheap talk in the trust game. *J Public Econom* 93: 1036–1044.
37. Greif A (1993) Contract enforceability and economic institutions in early trade: the Maghribi traders’ coalition. *Amer Econ Rev* 83: 525–548.

38. Greif A (2006) Institutions and the path to the modern economy. Cambridge: Cambridge University Press.
39. Diekmann A, Przepiorka W (2005) The evolution of trust and reputation: results from simulation experiments. Third ESSA Conference : 1–7.
40. Resnick P, Zeckhauser R, Swanson J, Lockwood K (2006) The value of reputation on ebay: A controlled experiment. *Exp Econ* 9: 79–101.
41. Bravo G, Tamburino L (2008) The evolution of trust in non-simultaneous exchange situations. *Rationality and Society* 20: 85–113.
42. Samuelson L, Zhang J (1992) Evolutionary stability in asymmetric games. *J Econom Theory* 57: 363–391.
43. Weibull JW (1995) Evolutionary Game Theory. Cambridge, MA: MIT Press.
44. Hofbauer J, Sigmund K (1998) Evolutionary Games and Population Dynamics. Cambridge, UK: Cambridge University Press.
45. Gintis H (2009) Game Theory Evolving, Second Edition. Princeton, NJ: Princeton University Press.
46. McNamara JM, Stephens PA, Dall SRX, Houston AI (2009) Evolution of trust and trustworthiness: social awareness favours personality differences. *Proc R Soc B* 276: 605–613.
47. Milinski M, Semmann D, Bakker TCM, Krambeck HJ (2001) Cooperation through indirect reciprocity: Image scoring or standing strategy? *Proc R Soc B* 268: 2495–2501.

## Figure captions

Figure 1: Schematic of the trust game.

Figure 2: Probability of Buy  $y_1^*$  in the cooperative equilibrium in the limit  $\mu \rightarrow 0$ . We set (a)  $\theta = 1$  and (b)  $\epsilon = 0.1$ .

Figure 3: The threshold value of  $r$  above which the cooperative equilibrium exists and is stable.

Figure 4: Boundary between the attractive basins of the cooperative and uncooperative equilibria. The points above the boundary are attracted to the cooperative equilibrium. We set  $\mu = 0.02$ ,  $\epsilon = 0.1$ ,  $r = 0.15$ , and  $\theta = 1$ .

Figure 5: Relative volume of the attractive basin of the cooperative equilibrium. We set  $\mu = 0.02$  and  $\epsilon = 0.1$ . Initially, (a) three and (b) four buyer's strategies are distributed according to the uniform density.

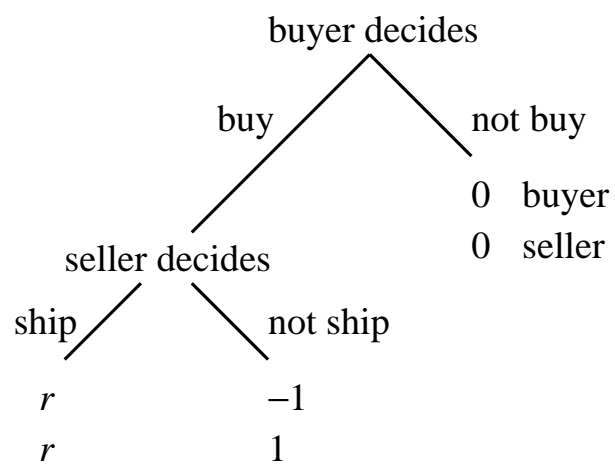


Figure 1

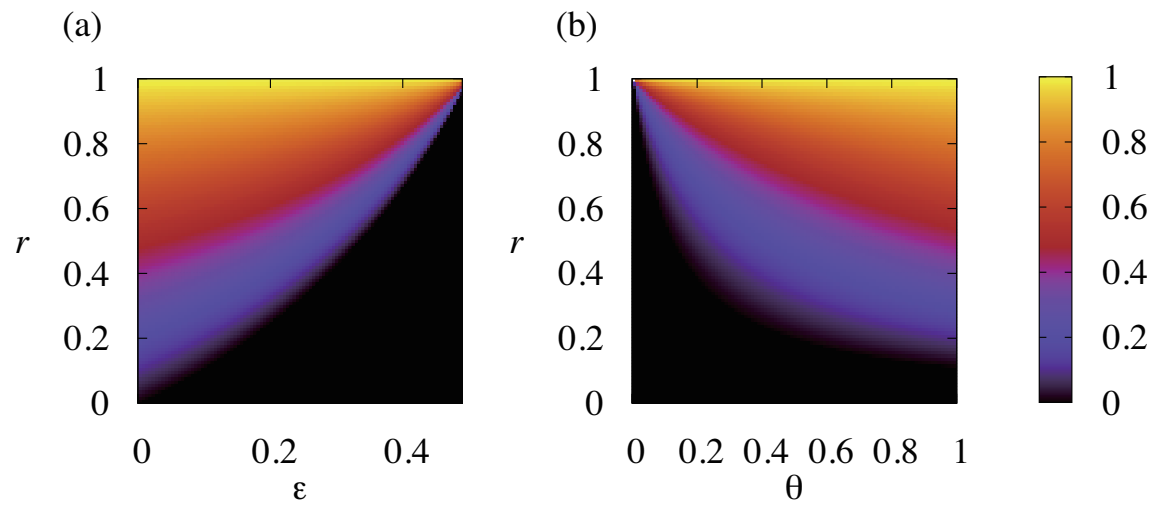


Figure 2

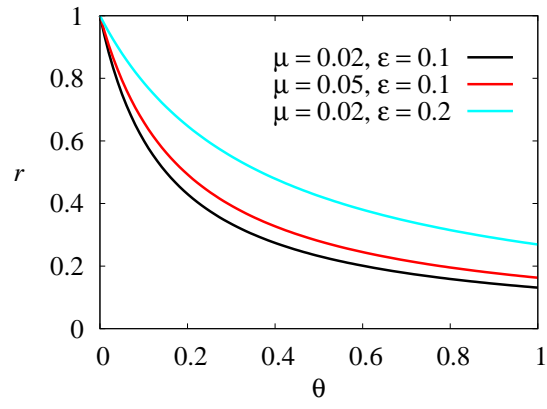


Figure 3

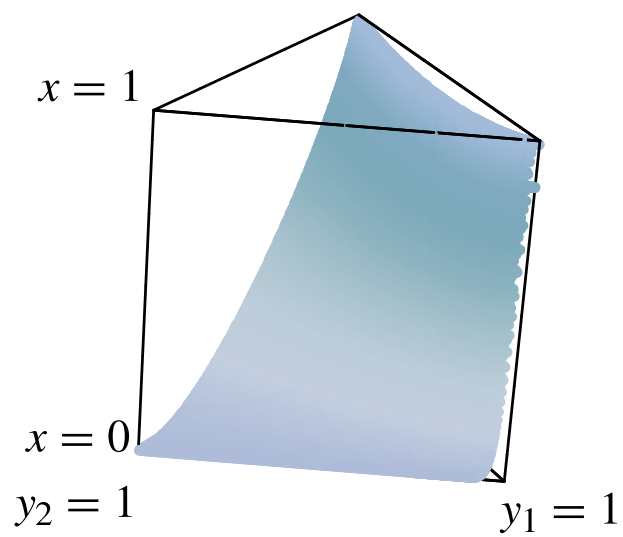


Figure 4

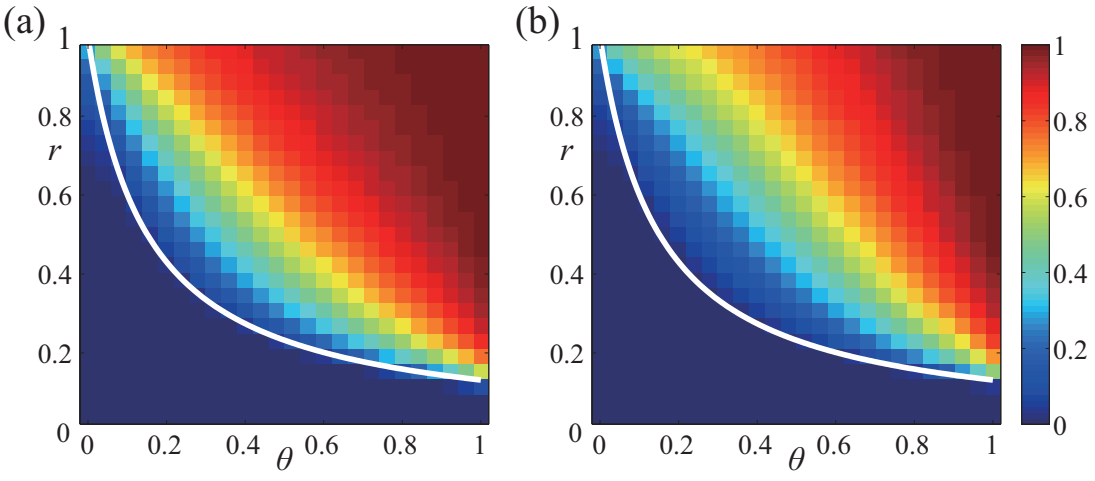


Figure 5